E-MMAD: Multimodal Advertising Caption Generation Based on Structured Information

Anonymous ACL submission

Abstract

With multimodal tasks increasingly getting popular in recent years, datasets with large scale and reliable authenticity are in urgent demand. Therefore, we present an e-commercial multimodal advertising dataset, E-MMAD, which contains 120 thousand valid data elaborately picked out from 1.3 million real product examples in both Chinese and English. Noticeably, it is one of the largest video captioning datasets in this field, in which each example has its product video (around 30 seconds), title, caption and structured information table that is observed to play a vital role in practice. We also introduce a novel task for vision-language research based on E-MMAD: e-commercial multimodal advertising caption generation, which requires to use aforementioned product multimodal information to generate textual advertisement. Accordingly, we propose a baseline method on the strength of structured information reasoning to solve the demand in reality on this dataset.

1 Introduction

Vision-and-Language has been drawing increasing attention from both computer vision and natural language processing communities, for there exists various multimodal information in real human life. As one of the most important tasks of vision-and-language (Uppal et al., 2021), multimodal text generation (Lin et al., 2021) aims to generate high-level text by fusing different modal effective information, such as video captioning (Lei et al., 2020a; Yang et al., 2019; Krishna et al., 2017).

However, there are few studies of multimodal text generation based on realistic multimodal data. One of the reasons is the lack of corresponding publicly available datasets, which can provide real-life multimodal information to help generate. Existing video-text generation datasets are mostly single modal input and are collected by manual batch-written templated descriptions such as MSR-VTT (Xu et al., 2016), Vatex (Wang et al., 2019). While in practice, information can also be divided into structured information and unstructured information. Humans tend to use richer structured information to generate appropriate text. This information can make the description rigorous and reliable. In this case, a large-scale and reliable dataset with structured information is in urgent demand.

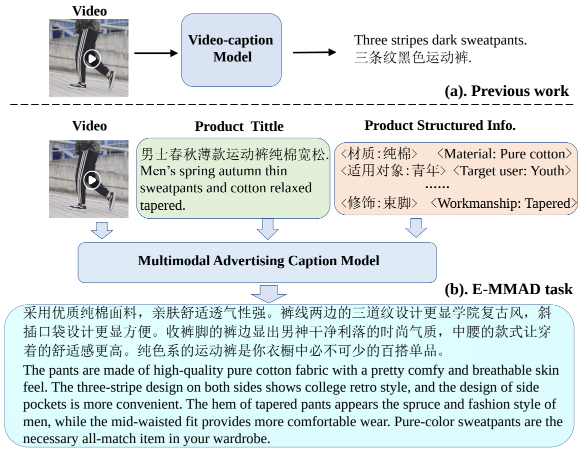

Figure 1: An illustration of our E-MMAD. The four different parts of our dataset, from top to bottom are product information(commodity displaying video, title, structured information) and commodity advertising description. The task of our model is to use the product information to generate corresponding advertising description. We add structured information to the original Video Caption to assist in generating a semantically richer caption

In this paper, we elaborately collect a large-scale e-commercial multimodal advertising dataset for multimodal text generation research, E-MMAD. To support in-depth research, we collect a rich set of product annotations. The E-MMAD dataset consists of 120,984 product instances in both Chinese and English, in which each instance has a product video, a title, structured information and a caption. Figure 1 illustrates a sample of our E-MMAD dataset. As is shown in Figure 1, E-commercial multimodal advertising generation task is typically more challenging than existing multimodal text generation, as the advertising description is vivid and information sources are abundant. More importantly, the caption needs to cover the information mentioned in the structured information table but missed in the video.

In response to the realistic demand for advertising generation, we propose the e-commercial multimodal advertising generation task and approach, which is qualified for better performance in generating approriate text. We propose the multimodal information fusion module and generation decoder module which make full use of the rich information. In addition, considering that various information words are often encountered in the process of model training and generalization, it will be difficult for the model to train. In the generalization process, since a considerable part of the nouns do not appear in the training, the caption quality generated by the model is not good enough. For example, when faced with unknown information including new brand names appearing in structured information, the model is not able to effectively identify and judge. So we propose Conceptualization Operations 4.1 to conceptualize complex and diverse information in real life as ontology. An ontology models generalized data, that is, we take into consideration general objects that have common properties and not specified individuals. Dataset and code will be available at our Website.

In summary, our contributions concentrate on the following three aspects:

(1) We collect a large-scale high-quality and reliable e-commercial multimodal advertising dataset. It is one of the largest video captioning datasets in this field. E-MMAD is collected from human real life scenes and carefully selected so that it is qualified to meet the needs of real life.

(2) We introduce a fresh task for vison-language research based on E-MMAD: e-commercial multimodal advertising generation, which requires to use the product multimodal information to generate textual advertisement.

(3) We propose a simple yet effective baseline method on the strength of structured information reasoning to solve the demand in reality on E-MMAD dataset.

2 Related Work

2.1 Multimodal video-text generation datasets

There are various datasets for multimodal video-text generation that cover a wide range of domains, such as movies (Rohrbach et al., 2015), cooking (Das et al., 2013; Zhou et al., 2018a), and Activities (Xu et al., 2016). MSR-VTT (Xu et al., 2016) is a widely-used dataset for video captioning, which has 10,000 videos from 257 activities and was collected in 2016. MSVD (Chen and Dolan, 2011) was collected in 2011, containing 1970 videos. ActivityNet (Caba Heilbron et al., 2015) has 20,000 videos but is used for Dense Video Captioning (Krishna et al., 2017), which means to describe multiple events in a video. TVR (Lei et al., 2020b) is collected from movie clips whose text is mainly character dialogue. Vatex (Wang et al., 2019) is a famous dataset released in 2019, whose caption is written by batch manpower. Compared with some mainstream datasets in Table 1, our dataset also provide an additional product structured information. We find that the advertising caption includes a lot of structured information in fact.

2.2 Video Captioning Approaches

Video caption/description is one of the important tasks in multimodal text generation. Early video caption methods are all based on templates (Mitchell et al., 2012; Krishnamoorthy et al., 2013). However, sentences made in this way tend to be rigid and stiff. The sequence-to-sequence model (Venugopalan et al., 2015) is a classic work, which includes an encoding phase and a decoding phase. After CNN extracts the image features of the video frames, an image feature is sent to the LSTM for encoding at each time step and text will be generated in the decoding stage. Some of the popular practices recently are based on data-driven (Zhang et al., 2021b) and transformer-based mechanisms (Yang et al., 2019; Zhou et al., 2018b; Lei et al., 2020a). MART (Lei et al., 2020a) can produce more coherent, non-repetitive, and relevant text to enhance the transformer architecture by using memory storage units. Vx2text (Lin et al., 2021) uses multimodal inputs for text generation. They use a backbone (Tran et al., 2018; Ghadiyaram et al., 2019) model to transform different modalities information to natural language and then the problem turns to natural language generation. Although good progress has been made by them, the original information of the modal is not fully utilized and integrated.

3 Datasets

In this section, we will introduce our dataset in detail, including the statistic analysis, collecting process, and comparison.

3.1 Data Collection

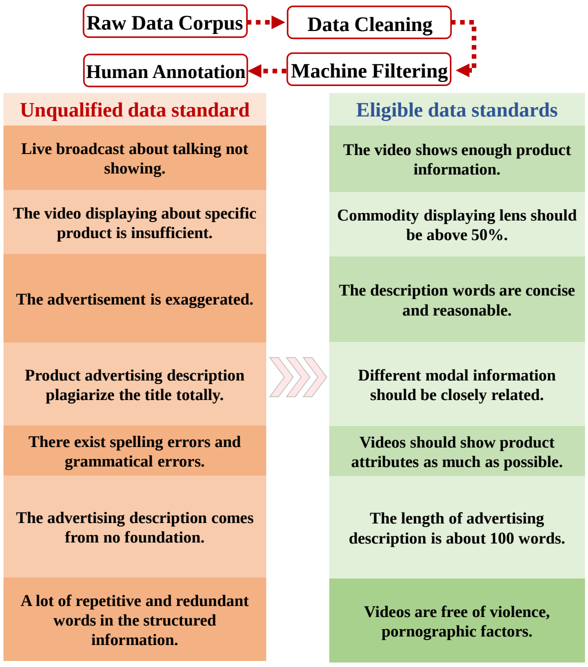

Figure 2: The process of creating a dataset, including cleaning, machine filtering, manual post-filtering, etc. and data specification of the dataset.

1) Dataset sources. Our dataset sources are the Chinese largest e-commerce website shopping platform (`www.taobao.com`), from which we have collected nearly 1.3 million commodity examples with structured information. It comprised more than 4,000 merchandise categories to guarantee the diversity of the dataset, such as clothes, furniture, office supplies, etc. The information of each commodity data sample includes structured information, commodity displaying video, title of product and commodity advertising description. Different from previous works (Wang et al., 2019; Xu et al., 2016; Chen and Dolan, 2011), the sources of datasets are derived from what merchants themselves numerously design and select, which comply with the standard rules of the authenticity of product advertisements and are supervised by false product advertising rules of *Taobao*. Specifically, videos visually display the commodity performance and application.

In addition, we fully consider ethical privacy issues to ensure that the dataset has no potential negative effects and legal issues (Gebru et al., 2018). All data is collected in *Taobao* shopping platform, which is a public platform for the general public. All information, even the characters in the video, is ensured to comply with Taobao laws including personal privacy, legal prohibitions, false information, protection of minors and women, and so on.

In consideration of data and ethics, we perform programmatic screening and manual cleaning again in accordance with the established data cleaning rules. Figure 2 shows our data collection process.

2) Data filtering. The intention for data filtering is to determine whether the product advertising description is closely related to the product displaying video, and whether the structured information of the product is in accordance with the composition of the product advertising description and ethical considerations. The product attributes structured information and product displaying video will be valid only if human being can write similar product advertising descriptions with them. We use programs to screen and judge at first, traversing the values of structured information. Our screening basis is the proportion of structured information words in the product advertising description. When the proportion is up to n words or more, the data will be reserved as valid data. After copywriters' continuous attempt to generate advertising descriptions with structured information words that account for different proportions, we finally determine the structured information with more than five words in the product advertising description as valid data and form 207,852 machine-screened data.
By virtue of this, we respectively test different groups of random data to formulate screening and judgment rules. Multiple copywriters tested and discussed to make the manual evaluation criterion several times. Finally, different testers sample 100 examples randomly according to the judgment rules of Figure 2, and the pass rate is mostly about 60%. In this case, we validate the manual screening rules and draw the conclusion that random subjective factors hardly have any influence. So far, the manual data screening and judging rules have been formed, as is shown in Figure 2.

3) Data annotation. We invited 25 professional advertising copywriters as data screening and annotation staff to conduct manual screening under the rules of Figure 2 and The Toronto Declaration . Manual screening of all data also ensures that each piece of data complies with the Toronto Declaration and *Taobao* laws to protect gender equality, racial equality, etc. In order to ensure the reliability of the data, we use the following two methods

Table 1: Comparison with other datasets. *Videos*, *Average Time*, *Caption Length*, *Classes* respectively represent the total number of videos in the dataset, the average video time in the dataset, the average length of the captions in the dataset and the number of instance types in the dataset. *Input Modality* indicates the input of the dataset, e.g. from Video to Text, Multimodal to Text. *Structure info.* means whether the dataset contains structured information. There are 3,876 keys of the structure information in E-MMAD dataset. en means English version dataset and zh means Chinese version dataset.

Datasets	#Videos	Average Time	Caption Length	#Classes	Input Modality	Structure Info.
MSR-VTT (Xu et al., 2016)	10,000	14.8s	9	257	Video	×
MSVD (Chen and Dolan, 2011)	1,970	9.0s	8	-	Video	×
TVR (Lei et al., 2020b)	21,800	9.0s	13	-	Video-query	×
VaTEX (en/zh) (Wang et al., 2019)	41,269	10.0s	15/13	600	Video	×
FFVD (zh) (Zhang et al., 2020)	32,763	27.7s	62	-	Video - Attribute	×
BFVD (zh) (Zhang et al., 2020)	43,166	11.7s	93	-	Video - Attribute	×
E-MMAD (en/zh)	**120,984**	**30.4s**	**97/67**	**4,863**	**Video - Title - Structure info.**	✓

to sample and verify: (1). Add verification steps. We will send back samples that have been annotated right answers to annotators from time to time to check their work quality. (2). Multiple people Choices. The data is sent to different people randomly. Only if the answers of all people are consistently passable, can this data be qualified. Finally, 120,984 valid data has been generated. Simultaneously, we also translate the filtered valid data into English so that both Chinese and English versions can be provided in the dataset. To ensure the quality of the English version, we use the WMT2019 Chinese-English translation champion, *Baidu machine translation*. We also monitor the translation quality in the manual screening section, such as random checking in batch translation, using text error correction to monitor retranslation, and back translation comparison.

After 25 people's diligent work of manual data labeling and cleaning, there are 120,984 valid data selected finally.

3.2 Dataset Analysis

Among the 207,852 data we send for annotation, there are 120,984 eligible samples passing the screening. We make an elaborate analysis on these valid data and the result is shown in Figure3. In addition to this, Figure 3 reveals the distribution of the product videos' duration and advertising descriptions.

By Table 1 comparison, we can find that our product advertising descriptions are not only at least twice longer than others, but also root in more vivid and realistic ones used in practice. The whole statistics about the structured information in our dataset is displayed in Figure 3 (d). What's more, there exist average 21 structured information words in each sample and 6.2 words of them are finally displayed in its product advertising description. The **(e)** shows the abundance of our datasets source classes.

3.3 Dataset Comparison

In Table 1, we make a comparison between our dataset and others from the following several perspectives: dataset scale, dataset diversity and dataset reliability.

1)Dataset scale: As shown in Table 1, the size of our E-MMAD is the largest multimodal dataset among those we have already known so far, with the longest video duration and text length, and the richest structured information in the dataset.

2)Dataset Diversity: In terms of types, our dataset consists of 4,863 categories. Our dataset is also available in Chinese and English two versions, to support multi-language research, which cannot be satisfied by a single language dataset. At the same time, our Chinese and English corpus is richer in vocabulary, which can generate more natural and diversified video descriptions.

3)Dataset Reliability: Compared with other manual batch-written descriptions(Wang et al., 2019) and mechanically generated data, our data annotation is derived from the real society. Each of them is an exclusive description genuinely written by corresponding store. Besides, the videos in our dataset are from the real product shooting scene, other than clips from Youtube or movies. We firmly believe that only resorting to reliable dataset, can we train models better. Therefore, we invest considerable amount of manpower and time in order to promote our dataset quality.

3.4 Dataset Significance

To the extent of our knowledge, the dataset we propose is the largest multi-modal dataset so far, and

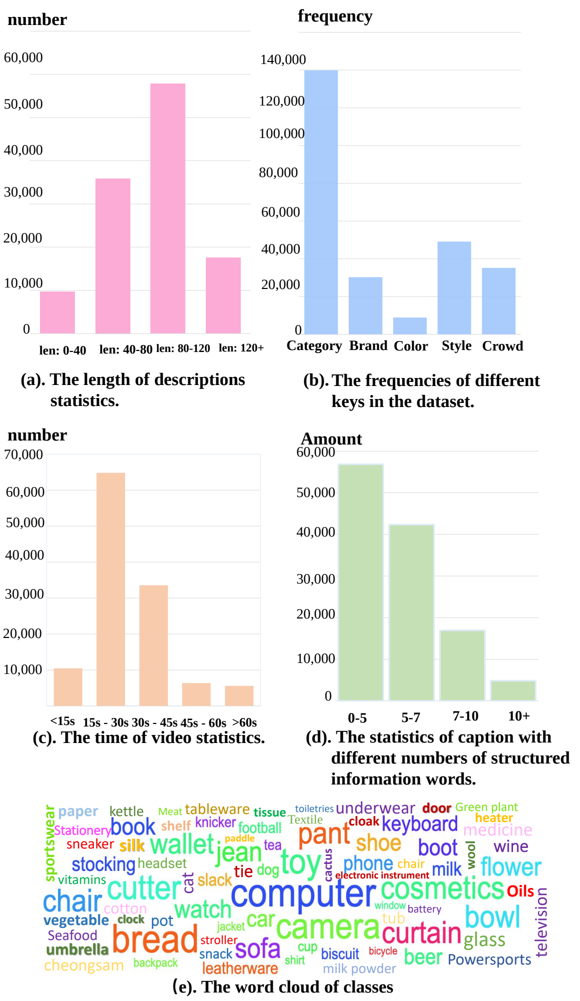

Figure 3: Statistics about the five different forms of data in our dataset. The data statistics are presented in terms of video, structured information, caption, and the main classes of the dataset contained, respectively.

the information involved is also the most diverse, which can better optimize and improve the performance of multi-modality models and promote their generalization ability to adapt to different scenarios in real world. For subsequent work, with the abundant and diverse information involved, our dataset can be dedicated to several multi-modality domain tasks, such as Video Retrieval (Lei et al., 2020b), Product Search (Chang et al., 2021) and so on. In our future work, we will build more versatile e-commerce datasets which can cover most tasks in this field based on this dataset.

4 Method

In this work, we present a novel approach called the Multi-modal Fusion and Generation algorithm as shown in Figure 4, which extracts feature representations from three sources: the product title, structured information(structured words) and the displaying video's frames and fuse them to generate captions. Faced with various information words, our model uses ontology, a method of conceptualizing information. That is to pre-process the various data, conceptualize and extract information from the complex information words to Key as highly conceptual network features. For the restoration of complex information in the generation phase, we only need to perform the inverse conceptualization operation at the end.

4.1 Conceptualization

During the training process, we pre-conceptualize the true product descriptions. The formula is as follows:

$$Values_{gr} = SW.values \bigcap GR.tokens; \quad (1)$$

$$k_{gr} \in SW.keys; \quad (2)$$

$$token_{gr} \rightarrow k_{gr},$$
$$\forall\, token_{gr} \in Values_{gr}. \quad (3)$$

In the generation process, the raw caption with conceptualized information generated by the model is de-conceptualized to obtain the final caption. The de-conceptualization is as follows:

$$Values_{rc} = SW.keys \bigcap RC.tokens; \quad (4)$$

$$v_{gr} \in SW.values; \quad (5)$$

$$rc\_token \rightarrow v_{gr},$$
$$\forall\, rc\_token \in Values_{rc}. \quad (6)$$

Among them, Equation 3/6, $A \rightarrow B$ means replacing token A with token B. $A \in C$ means token A is an element of set C. $GR.tokens$ and $RC.tokens$ are the sets of corresponding n-gram phrases in ground truth and raw caption, respectively. $SW.values$ and $SW.keys$ respectively correspond to the sets of keys and values in the structured information. In terms of the model input, the ontology of the structured information part is conceptual value words. An ontology models generalized data, that is, we take into consideration general objects that have common properties and not specified individuals. By this, the 3,876 types of Keys represent the various information words as the highly conceptual network feature input. We also reference the title as the basis to determine the priority position of each key according to the order in which the structured information appears in the title.

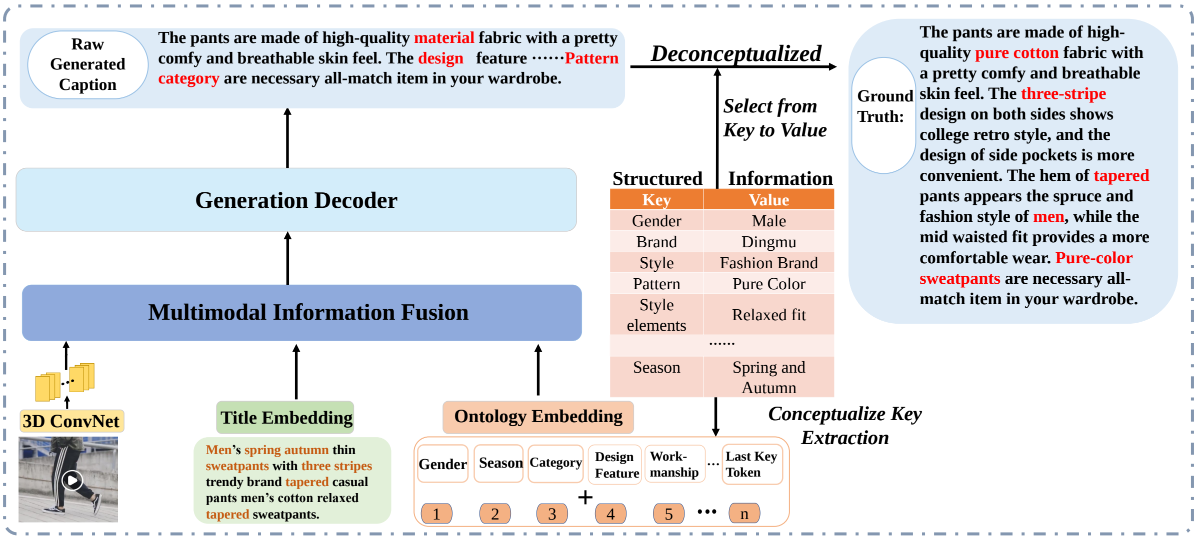

Figure 4: The overall architecture of our model, which contains three main parts: the representation for multimodal information, the multimodal fusion module based on self-attention and the generation decoder module on the basis of (Radford et al., 2019). According to the Key-Value, the used Structure information words are conceptualized as ontology to face the various words such as assorted brands in real life.

4.2 Representation

Textual Information. Given a product title as a list of K words, conceptualized product attributes as a list of N keys, we embed these words and keys into the corresponding sequence of d-dimensional feature vectors using trainable embeddings (Zhang et al., 2021a; Devlin et al., 2018). In addition, since the keys of structured information are prioritized, we use *position embedding* to represent the priority position of the keys.

Visual Information. Given a sequence of video frames/clips of length S, we feed it into pre-trained 3D CNNs(Ji et al., 2012) to obtain visual features $V = \{v_1, v_2, \ldots, v_K\} \in \mathbb{R}^{S \times d_v}$, which are further encoded to compact representations $R \in \mathbb{R}^{S \times d}$, which have the same dimension as the representation of textual information via a *Visual Embedding Layer*. The *Visual Embedding Layer* can be formalized as following:

$$f_{VEL}(v) = \mathrm{BN}(g \circ \bar{v} + (1-g) \circ \hat{v}); \quad (7)$$

$$\bar{v} = W_1 v^\top; \quad (8)$$

$$\hat{v} = \tanh\left(W_2 \bar{v}\right); \quad (9)$$

$$g = \sigma\left(W_3 \bar{v}\right). \quad (10)$$

BN denotes batch normalization, $\circ$ is the element-wise product, σ means sigmoid function, $W_1 \in \mathbb{R}^{d \times d_v}$ and $\{W_2, W_3\} \in \mathbb{R}^{d \times d}$ are learnable weights.

4.3 Multimodal Fusion

After embedding all information from each modality as vectors in the d-dimensional joint embedding space, we use a stack of L transformer layers with a hidden dimension of d to fuse the multi-modal information consisting of a list of all $K + N + S$ modalities from $\{v_S^{\text{frames}}\}$ $\{v_K^{\text{words}}\}$ and $\left\{v_N^{\text{keys}}\right\}$. Through the self-attention mechanism in transformer, we can model inter- and intra- modality context. The output from our Multimodal Information Fusion and Reinforcement module is a list of d-dimensional feature vectors for entities in each modality, which can be seen as their interrelated embedding in multimodal context. In this work, the parameters chosen for our the module are consistent with the parameters of *BERT-base* (L=12, H=768, A=12), where L, H, A represents the number of layers, the hidden size, and the number of self-attention heads respectively.

4.4 Generation Decoder

Our model's decoder is a left-to-right Transformer decoder, which is similar to the model architecture of (Chen et al., 2019; Radford et al., 2018). The decoder accesses multimodal fusion outputs at each layer with a multi-head attention (Vaswani et al., 2017). Specifically, the decoder applies a multi-headed self-attention over the caption textual feature. After that, the position-wise feed forward layer was used to produce a distribution probability of each generation tokens for the final generated

Table 2: Performance (%) comparison with our proposed model and others. The NACF + multi-input means that we concat the structured information and title with video feature directly as input. On the premise of fair comparison, the following methods are relatively classic and available, which are applicable on E-MMAD by our objective attempts.

Version	Input	Method	Bleu1	Bleu2	Bleu3	Bleu4	Rouge_L	CIDEr
en	Text	NLG (Chen et al., 2019)	13.6	6.8	3.1	1.9	13.0	10.1
	Video	NACF (Yang et al., 2019)	18.9	7.9	3.9	2.2	15.3	14.8
	Multimodal	NACF + multi-input	20.0	8.5	4.3	2.4	17.8	18.5
		TVC (Lei et al., 2020b)	21.3	12.4	6.2	3.7	19.3	22.5
		Ours (en)	**25.0**	**16.6**	**9.6**	**7.2**	**25.3**	**29.1**
zh-CN	Text	CPM (zh) (Zhang et al., 2021a)	7.9	4.6	1.1	0.5	7.2	8.3
	Multimodal	ours (zh)	**11.6**	**6.5**	**4.4**	**2.2**	**12.5**	**15.3**

caption. There is a description of part of the formula for the decoder module:

$$h_0 = V^{\text{cap}} \cdot W_t + PE \cdot W_p; \tag{11}$$

$$h_l = \text{ Trans\_Block } (h_{l-1}); \tag{12}$$

$$P(w) = \text{Softmax}\left(h_n W_e^T\right); \tag{13}$$

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{\text{model}}}\right); \tag{14}$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{\text{model}}}\right); \tag{15}$$

where $V^{cap} = \{v_1, v_2, \ldots, v_x\}$ is the textual vector of caption, n is the number of layers, $\forall l \in [1, n]$, and W_t, W_p is the learnabale weight for caption embedding feature and position encoding respectively. $Trans\_Block$ represents a block of the decoder in the Transformer (Vaswani et al., 2017). We refer to (Radford et al., 2019) as the model decoder architecture.

5 Experiments

In this section, we will show a series of experiments of our proposed model on E-MMAD, including ablation studies, comparison experiments and state-of-the-art video caption methods and human evaluation.

5.1 Implementation Details

All the experiments are conducted on Nvidia TitanX GPU. The proposed model is implemented with PyTorch. For the representations of videos, we follow (Yang et al., 2019) for fairness and opt for the same type, first extract 3D features with 2048 dimensions, 2048-D image features from ResNet-101 (Hara et al., 2017) pre-trained on ImageNet dataset. For generation decoder, we use <*sep*> to separate the input from the ground truth of caption. We adopt diverse automatic evaluation metrics to compare with other model: BLEU (Papineni et al., 2002), Rouge-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). It is worth noticing that the focus of the CIDEr evaluation metric is on whether the generated caption captures the major information or not. Since the major information captured by each model is different, the key information component of the generated caption will not be the same, but it is cognitive at the semantic level, so the CIDEr evaluation metric will have a relatively large fluctuation. Our model introduces structured information so that the generated caption can include most of the major information. Therefore, the caption generated by our model can achieve significant results in the evaluation index of CIDEr.

5.2 Comparison with Other Approaches

During the comparison experiments, we uniformly divided the Chinese and English versions of our dataset into training set, validation set and test set in the ratio of 6:2:2 for training and testing. Since the current mainstream models do not use multi-modal data for captioning, we use unimodal data for captioning on some classic and available methods, such as video caption, nlg, etc. For the sake of fairness of comparison, we simply modify the input part of the above experimental model to accommodate multimodal data. As we can see from Table 2, the comparison of the results before and after the model modification shows that multimodal data can substantially improve text generation tasks. It indicates that multimodal information indeed helps captioning by modal information between the mutual enhancement. As shown in Table 2 our algorithm achieves a better performance than other methods because our model makes better use of multimodal data in the means of fusing different modalities and structured information to reason.

5.3 Ablation studies

Multimodal Input. We perform ablation studies based on changing the input components of our proposed model as a way to validate the importance of our proposed dataset containing structured information. As shown in Table 3, we analyze the gap between the generated caption of the model and the real commodity advertising description in the absence of partial information. As we can see, the absence of any of the three input components significantly degrades the final generated caption result. From our analysis of the generated caption, we can conclude that: 1) the lack of structured information will make the generated caption less informative, rigorous and reliable.

Table 3: Performance comparison with our proposed model by masking different parts of input and only using the remainder as input. Here "Title", "SI" and "Video" indicates commodity title, structured information and commodity displaying video respectively.

Input	Bleu1	Bleu2	Bleu3	Bleu4	Rouge_L	CIDEr
SI & Video	22.8	14.8	6.9	5.5	22.2	25.3
Title & Video	19.5	9.4	4.5	3.1	16.4	15.7
Video	15.9	6.4	3.4	2.1	15	13.2
Title & SI	22.0	13.8	5.8	4.9	20.6	23.7

2) The lack of a commodity title or displaying video will impair the foundation of generated text. In addition, the structured information is like a knowledge base, which can promote inference and judgment to generate appropriate caption.

Conceptual Operation. Considering that writing product descriptions in real life often involves a great number of unfamiliar words, which makes it hard for the model to identify and remember its feature when facing a new word, such as new brand name. The predecessor's approach tend to use as much corpus and large model parameters as possible, which brings huge difficulties to natural language generation. In this case, we proposed the Conceptualization operation. As shown in Table 4 , we conduct ablation experiments about Conceptualization on the Chinese and English datasets. As for models without conceptual operations, we use unconceptualized captions as the ground truth to train. We directly input unordered structured words for the input of the model. Experiments have proved that the Conceptualization operation can indeed bring a significant effect improvement, because this method can conceptualize and extract information from complex information in the dataset, and thus highly conceptualize network features. We expect this discovery to inspire the community.

Table 4: Performance comparison of whether our proposed model has conceptual operations (CO).

Operation	Bleu1	Bleu2	Bleu3	Bleu4	Rouge_L	CIDEr
ours w/o CO (en)	23.8	15.4	8.1	6.4	24.2	27.3
ours w/o CO (zh)	9.9	5.5	2.8	1.5	10.1	12.4
ours w/ CO (en)	**25.0**	**16.6**	**9.6**	**7.2**	**25.3**	**29.1**
ours w/ CO (zh)	**11.6**	**6.5**	**4.4**	**2.2**	**12.5**	**15.3**

5.4 Human Assessment

It is well-known that the human evaluation metrics for video captioning are required due to the inaccurate evaluation by automatic metrics. We especially focus on advertising generation, which depend on human aesthetics. So we invite the people involved in the data annotation and new advertising slogan designers to conduct the human evaluation. We select 200 samples from the test dataset and each evaluator evaluate each of these 200 samples to reflect the performance of our model by rating whether the caption generated by our model can be used as a description of the product. As the result shows in Table 5, the caption generated by our model has a certain degree of pass rating, whose results can be approbated by people. Therefore, this is also acceptable that our experiments on Table 2 did not achieve high scores for mechanical evaluation indicators.

Table 5: The results of the human evaluation, reflecting the proportion of the 200 examples where the model generated caption could be used as a product description that describes the reasonableness of the generated caption. Annotators are from the dataset annotation and persons are from the frequent online shopping masses.

	Annotator 1	Annotator 2	Annotator 3	Person 1	Person 2	Person 3
Pass	42%	44%	43%	48%	56%	53%

6 Conclusion and Future Work

This research sets out to provide an e-commercial multimodal advertising dataset, E-MMAD, which is one of the largest video captioning datasets in this field. Based on E-MMAD, we also present a novel task: e-commercial multimodal advertising generation, and propose a baseline method on the strength of structured information reasoning to solve the realistic demand. We hope the release of our E-MMAD would facilitate the development of multimodal generation problems. However, there still exist limitations about our dataset and method that should be acknowledged. Moving forward, we are planning to extend E-MMAD to better performance and more diversified tasks by exploring new model structures, fine-grained and so on.

References

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Wei-Cheng Chang, Daniel Jiang, Hsiang-Fu Yu, Choon Hui Teo, Jiong Zhang, Kai Zhong, Kedarnath Kolluri, Qie Hu, Nikhil Shandilya, Vyacheslav Ievgrafov, et al. 2021. Extreme multi-label learning for semantic matching in product search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2643–2651.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2019. Few-shot nlg with pre-trained language model. *arXiv preprint arXiv:1904.09521*.

Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. 2019. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12046–12055.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Learning spatio-temporal features with 3d residual networks for action recognition. *arXiv preprint*, arXiv:1708.07632.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3d convolutional neural networks for human action recognition. volume 35, pages 221–231. IEEE.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, and Sergio Guadarrama. 2013. Generating natural-language video descriptions using text-mined knowledge. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. 2020a. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402*.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020b. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. 2021. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7005–7015.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alexander Berg, Tamara Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. 2021. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Bang Yang, Yuexian Zou, Fenglin Liu, and Can Zhang. 2019. Non-autoregressive coarse-to-fine video captioning. *arXiv preprint arXiv:1911.12018*.

Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020. Poet: Product-oriented video captioner for e-commerce. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1292–1301.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, et al. 2021a. Cpm: A large-scale generative chinese pre-trained language model. *AI Open*, 2:93–99.

Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. 2021b. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9837–9846.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748.

A Appendix Ablation Results Tables

Source Link: https://github.com/E-MMAD/E-MMAD